SUBBAND BASED CLASSIFICATION OF SPEECH UNDER STRESS

*Ruhi Sarikaya** and *John N. Gowdy*
Digital Speech and Audio Processing Laboratory
Department of Electrical and Computer Engineering
Clemson University, Clemson, SC 29634
ruhi@ee.duke.edu john.gowdy@ces.clemson.edu

ABSTRACT

This study proposes a new set of feature parameters based on subband analysis of the speech signal for classification of speech under stress. The new speech features are Scale Energy (SE), Autocorrelation-Scale-Energy (ACSE), Subband based cepstral parameters (SC), and Autocorrelation-SC (ACSC). The parameters' ability to capture different stress types is compared to widely used Mel-scale cepstrum based representations: Mel-frequency cepstral coefficents (MFCC) and Autocorrelation-Mel-scale (AC-Mel). Next, a feedforward neural network is formulated for speaker-dependent stress classification of 10 stress conditions: Angry, Clear, Cond50/70, Fast, Loud, Lombard, Neutral, Question, Slow, and Soft. The classification algorithm is evaluated using a previously established stressed speech database (SUSAS)[4]. Subband based features are shown to achieve **+7.3**% and **+9.1**% increase in the classification rates over the MFCC based parameters for ungrouped and grouped stress closed vocabulary test scenarios respectively. Moreover the average scores across the simulations of new features are **+8.6**% and **+13.6**% higher than MFCC based features for the ungrouped and grouped stress test scenarious respectively.

1. INTRODUCTION

Stress is defined as perceptually induced deviation in the production of speech from that of the normal production of speech [3]. It is known that stress-based variations in speech production can be substantial and will hence deteriorate the performance of speech processing applications [1, 6, 3, 5]. If the knowledge of stress and its type could be determined then this *extra* information could be incorporated into a speech recognition or coding system to improve system performance [3]. One real-time application of stress detection and classification is for a metropolitan emergency telephone system in which such a system can be used to direct the emotional telephone calls to a priority operator [3]. Such a system could also be used for aircraft voice communication monitoring.

The manner in which stress manifests itself in speech signals has been studied by many researchers. These studies investigated the effects of stress in speech in different domains in order to derive reliable acoustic features for stress classification [1, 6, 3]. For example in [6] glottal waveforms were modeled and parametrized for the purpose of investigating the variations in glottal excitation across stress conditions. In [9], the acoustic-phonetic differences among various stress conditions were considered in the following parameter domains: energy bands, spectral center of gravity, spectral tilt, pitch, formant locations, and duration. They observed the most reliable trends in the energy migration in frequency domain. It was also noted that for Loud and Lombard speech, the speakers typically move additional energy into low to mid-bands which is the frequency range of greatest sensitivity of the human auditory system. We also observed the same phenomenon for stressed speech which motivated us to be able to formulate a good representation of energy and energy migrations among subbands. In [3], Mel-scale cepstrum based parameters for stress classification were proposed. Although Mel-scale analysis incorporates the properties of the human auditory system into analysis, performing Mel-scale warping on the vocal tract spectra inherently ignores the excitation spectra which is the major relayer of stress. It is known [6] that the stress information in the speech signal is mainly carried by the excitation rather than the vocal tract in the linear modeling of speech. Therefore, an effective feature set must address the issues outlined above.

We propose a new set of features based on wavelet analysis or equivalently the multirate subband analysis of stressed speech. The resulting parameters have special features which make them useful in this application. First, the subband decomposition provides a local spectral representation for stressed speech. In the short-time Fourier transform (STFT) of the speech signal, time-frequency resolution is fixed once a window has been chosen for that speech segment. However, the multiresolution nature of the wavelet analysis permits the observation of local spectral variations within the windowed segment. Second, perceptual division of the frequency axis can be obtained by appropriate choice of the wavelet packet tree to account for the human auditory property. Finally, better frequency localization compared to STFT based methods can be achieved in order to model energy migrations among subbands by choosing filters which have maximum vanishing moments.

2. FEATURE EXTRACTION

2.1 Subband Decomposition via Wavelet Packets

Although a detailed discussion of wavelets, wavelet transform and wavelet packets is beyond the scope of this paper,

*R. Sarikaya was with Clemson University when this work was performed. He is now with Robust Speech Processing Laboratory, Duke University, Durham, NC 27708-0291

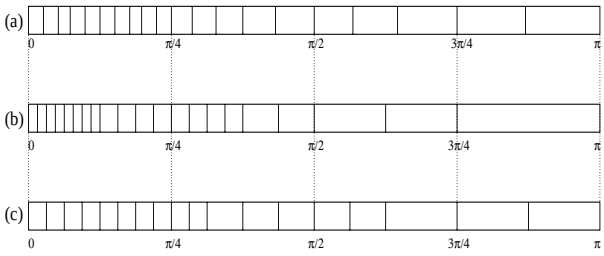

Figure 1: Frequency divisions (a)Mel-Scale, (b) 21-subband, (c) 18-subband wavelet packet trees.

we refer readers wishing to see a complete discussion presented in [8]. The Wavelet Transform is defined as the inner product of a signal $x(t)$ with a collection of wavelet functions $\psi_{a,b}(t)$ in which these wavelet functions are scaled (by a) and translated (by b) versions of the prototype wavelet; $\psi(t)$.

$$\psi_{a,b}(t) = \psi\left(\frac{t-a}{a}\right) \tag{1}$$

$$W_\psi x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right) dt \tag{2}$$

Discrete time implementation of wavelets and wavelet packets are based on iteration of two channel perfect reconstruction filterbanks. Contrary to wavelets which are obtained through iterations on the low pass branch, the filterbank tree can be iterated on either branch at any level resulting in a tree structured filterbank which we call a wavelet packet filterbank tree. The resultant transform creates a division of the frequency domain that represents the signal optimally with respect to the applied metric. Because of the nature of the analysis in the frequency domain, it is also called subband decomposition where subbands are determined by a wavelet packet filterbank tree.

In this study two wavelet packet trees are used with the goal of perceptual division of the frequency axis. The resulting wavelet packet divisions and the Mel-scale division are shown in the Figure 1.

2.2 Subband Based Feature Extraction Procedure

The speech signal is sampled at 8kHz and divided into overlapping frames of equal size. For word tokens, energy based end-point detection is performed. For phoneme tokens voiced/unvoiced classification is performed by using the Simple Inverse Filter Tracking (SIFT) algorithm [10]. Then, voiced phones are extracted from the stressed words to be used in classification. Two frame lengths have been used for feature extraction to explore the effect of frame lengths on classification results. Without regarding the duration of the word or phoneme token, 40 frames are obtained with a variable skip rate. For 16 msec frame size the minimum degrees of overlap for phonemes and words is 43% and 3%, respectively. For 24 msec frame size the corresponding quantities are above 56% and 37% for phonemes and words, respectively.

The steps of feature extraction are explained in the block diagram given in Figure 2. After segmentation, each frame of speech is decomposed into subband signals by using the perceptual wavelet packet transform which is implemented by cascaded filterbanks along the wavelet packet tree. The output of this transformation is a set of subband signals or equivalently transform coefficients.

2.3 Scale Energy Parameters (SE)

The energy of the subsignals for each subband is computed and then scaled by the number of transform coefficients in that subband. This operation results in an energy vector which is normalized by the total energy in that frame to give the SE parameters:

$$S_i^{(k)} = \frac{\sum_{mei}[(W_\psi x)(i),m)]^2}{N_i} \tag{3}$$

$$\underline{SE}^{(k)} = \frac{\underline{S}^{(k)}}{|\underline{S}^{(k)}|} \tag{4}$$

$W_\psi x$: wavelet packet transform of x,
k : frame number,
i : subband number ($i = 1, 2...L$),
N_i : number of coefficients in the i^{th} subband,
n : spans all available frames, and
$\underline{SE}$: parameter vector for each frame.

SE parameters represent the distribution of energy among various frequency bands for a given frame. Since we use the orthogonal filters corresponding to Daubechies's orthogonal wavelets in the wavelet packet transform, energy is preserved in the transformation.

2.4 Autocorrelation-SE Parameters (ACSE)

ACSE parameters are motivated by AC-Mel parameters [3]. First proposed in [1], ACSE parameters are given by,

$$ACSE_{(i)}^{(l)}(k) = \frac{\sum_{n=k}^{k+T}[SE^{(n)}(i)SE^{(n+l)}(i)]^2}{argmax_j(ACSE_{(i)}^{(l)}(j))} \tag{5}$$

where k is the frame index for ACSE parameters, T is the correlation window length, j is an index which spans all correlation coefficients in a given scale and l is the correlation lag over frames. When $l = 0$, ACSE models the normalized power in subband i. It is both a measure of the frame-to-frame correlation variation of SE parameters and the relative change in subband energies due to stress. T and l are chosen 6 and 1, respectively, as suggested in [3] for the entire simulations.

2.5 Subband based Cepstral Parameters (SC)

SC parameters are derived from SE parameters by applying the following transformation:

$$SC(k) = \sum_{i=1}^{L} S_i^{p_i} cos\left(\frac{k(i-0.5)}{L}\pi\right), \quad k = 1, 2, ...k' \tag{6}$$

where k' is the number of SC parameters, p_l is the root value of the l-th frequency band and L is the total number of frequency bands. A slightly different version of these features was proposed in [11] for recognition in noisy environments. Because of the similarity to root-cepstral [7] analysis they are named as subband based cepstral analysis. The advantage of this representation is that the frequency bands can be emphasized or deemphasized by appropriate choice of p_l for each subband. Since we are processing noise free stressed speech we have used uniform $p_l = 0.375$ [7] for all subbands.

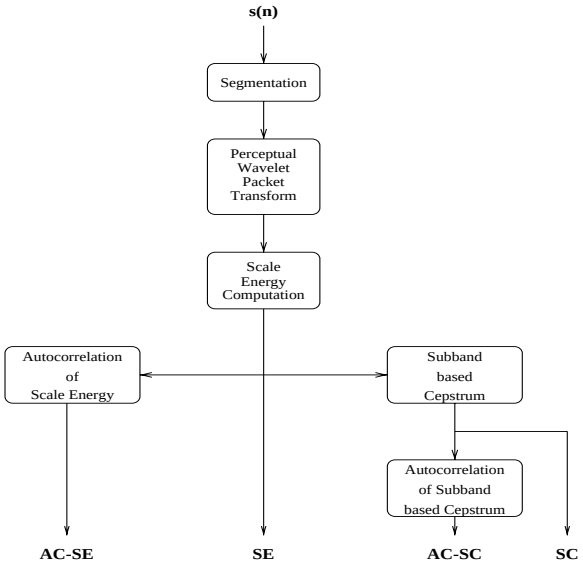

Figure 1: Block diagram for subbands based feature extraction procudure.

2.6 Autocorrelation-SC Parameters (ACSC)

ACSC has a mathematical expression similar to ACSE and has a dual interpretation. The relation between SC and ACSC is same as the relation between SE and ACSE parameters. They can be interpreted as the frame-to-frame correlation of SC parameters and relative changes in subband energy which models energy migration among subbands.

2.7 MFCC and AC-Mel

A detailed performance analysis of MFCC based parameters has been performed in [3] in the context of stress classification. In [3], four MFCC based parameters were proposed and MFCC and AC-Mel were found to outperform the other proposed features. In order to compare the performance of the subband based features with the MFCC based features, we duplicate part of their simulations in this study. MFCC parameters are obtained in the frequency domain with critical band windowing of the speech spectrum. AC-Mel parameters are derived from from MFCC parameters the same way ACSE parameters obtained from SE parameters.

3. STRESSED SPEECH CLASSIFICATION

3.1. Stress Classification

One of the two sets of 5 word speech corpus obtained from SUSAS [4] is used in the training phase while the other in the testing phase. The training set is composed of five words: *break, east, freeze, help, steer* spoken under 10 stress conditions. The testing set is composed of same set of words but spoken at a second time by the same speaker. We used the same vocabulary set for testing in order to simulate the intra speaker variability of stress conditions across same set of words for a given speaker. In order to provide a fair comparison, the data set chosen for training and testing here is the same data set used in [3].

3.2. Neural Network Classifier

A feedforward multi-layer-perceptron (MLP) architecture with backpropagation training method is formulated as the stress classifier. The size of the input layer is dependent on the feature vector size which varies between 612 to 840 (40 frames x 21 subbands) for each token. Although, there is not a clear rule for choosing the number of layers and the number of neurons in a given layer, Kolmogorov's Mapping Neural Network Existence Theorem [12] indicates that a four layer network having d processing units in the first layer and $2d+1$ units in the next layer is sufficient to map an arbitrary continuous function. However, since the proof of the theorem is not constructive, it is not possible to know how to determine the key quantities of the transfer functions. The theorem simply tells us that such a network must exist. Therefore, we set the size of the first hidden layer and the second hidden layers to 1200 and 300 neurons, respectively. The output represents the stress classes. The stress condition which receives highest score in the output layer is selected as the winner. The classifier used in our simulations is not optimal by any means. However the reader should note that the primary goal of this work is not to find the best classifier but rather to compare the performance of the subband based features with MFCC based features for a given classifier.

4. SIMULATIONS AND DISCUSSIONS

4.1 Simulation Scenarios

In order to optimize the performance of the features under consideration frame length, speech token type, and number of subbands are treated as variables. Wavelet packet trees with 18 and 21 subbands approximating Mel-scale frequency division were used to generate features. Frames of size 16 msec and 24 msec are used to investigate the effect of frame length on the performance of the generated features. Additionally, the features are derived from the word and the chosen voiced phoneme within the word with the goal of investigating effects of stress on the phoneme level and isolated word level. The tests were conducted for ungrouped-stress and grouped-stress closed-vocabulary cases. Consequently, 16 simulations were conducted for each of the six parameters sets.

4.2. Discussion

In ungrouped stress classification, *Angry, Loud, Lombard, Cond50/70*[1] and *Question* stress styles consistently obtained relatively high classification rates in every parameter domain whereas *Neutral, Fast* and *Slow* styles had a very low classification rates. These 3 stress conditions are confused during the classification. These styles have very similar spectral distributions. The only distingushing feature among these styles is phone duration. However, we obtained 40 frames from each token regardless of its duration. This is essentially the linear time warping of the data. Normalizing the duration accounts for the low classification rates for these styles. The results shown in the Table 1 are the average of results obtained by using different number of subbands (parameters), different frame lengths and different token types. The reader wishing to see individual simulation results for each scenario should look at [2]. While the overall classification rates of MFCC based parameters are around 45%, subband based parameters achieved higher classification rates.

[1] Represents stressed speech produced during computer workload tasks.

Table 1: Ungrouped Stress Classification

AVERAGE STRESS CLASSIFICATION SCORES (%)						
STRESS CLASS	MEL-CEPSTRAL		SUBBAND FEATURES			
	MFCC	AC-MEL	SE	SC	ACSE	ACSC
Angry	90.0	65.0	57.5	95.0	57.5	57.5
Clear	7.5	17.5	15.0	22.5	15.0	17.5
Cond50/70	66.2	55.0	72.5	71.2	67.5	70.0
Fast	5.0	32.5	10.0	15.0	35.0	37.5
Loud	82.5	80.0	45.0	82.5	85.0	47.5
Lombard	60.0	52.5	75.0	90.0	57.5	70.0
Neutral	0.0	12.5	5.0	10.0	10.0	10.0
Question	55.0	70.0	80.0	90.0	97.5	85.0
Slow	0.0	22.5	2.5	7.5	32.5	7.5
Soft	50.0	30.0	85.0	95.0	47.5	57.5
OVERALL	44.3	45.4	47.0	59.1	52.5	48.4

Table 2: Grouped Stress Classification

AVERAGE STRESS CLASSIFICATION SCORES (%)						
STRESS CLASS	MEL-CEPSTRAL		SUBBAND FEATURES			
	MFCC	AC-MEL	SE	SC	ACSE	ACSC
G1	100.0	93.7	67.5	96.3	76.3	78.8
G2	90.6	73.1	95.0	95.6	90.0	81.8
G3	10.0	32.5	10.0	7.5	20.0	22.5
G4	45.0	60.0	55.0	87.5	87.5	75.0
G5	5.0	20.0	2.5	2.5	12.5	2.5
G6	0.0	12.5	10.0	12.5	27.5	15.0
G7	47.5	50.0	72.5	72.5	57.5	62.5
OVERALL	61.4	59.5	60.4	70.0	65.7	60.2

In particular, SC parameter received 59.1% which is 13.6% higher than MFCC based parameters on average.

In order to investigate whether improved classification scores could be obtained, some of the stress conditions are combined into groups. *Angry* and *Loud* styles are combined in *(G1)* and *Cond50/70, Neutral* and *Soft* styles are grouped under *(G2)*. Each of the other stress styles are treated as a separate group of its own :*Fast (G3), Question (G4), Slow (G5), Clear (G6)*, and *Lombard (G7)*. Although this grouping may not be optimal for our classifier, it was the same grouping proposed in [3]. Stress grouping improved the performance of each parameters set between 11% to 17%. *G1* and *G2* received the highest scores across all parameters. *G3, G5* and *G6* had poor scores because of the duration normalization required by our classifier. The results given in Table 2 indicate that SE and ACSC performed as well as MFCC based parameters whereas ACSE and SC received higher scores. According to results of these simulations SC consistently yielded the highest classification rates.

The results given in Table 1 and 2 are the average scores across simulations. The best scores of each parameter set on individual simulations are given in Table 3. In individual simulations, subband parameters achieved +7.3% and +9.1% higher classification scores than MFCC based features in ungrouped and grouped stress classification. The simulation results for ungrouped-stress indicate that combination of 21 subbands, word tokens and 24 msec frame size gives the highest rates for all parameters. For grouped-stress simulations the optimum frame size and the number of subbands were again 24 msec and 21 subbands with the exception of ACSC which obtained highest score when 18 subbands were used. An interesting observation is that all autocorrelation parameters (AC-Mel, ACSE, ACSC) obtained the highest rates when word tokens instead of phoneme tokens were used whereas the non-autocorrelation parameters

Table 3: Best Scores of Each Feature

BEST STRESS CLASSIFICATION SCORES (%)						
TESTING CASES	MEL-CEPSTRAL		SUBBAND FEATURES			
	MFCC	AC-MEL	SE	SC	ACSE	ACSC
Ungrouped	50.9	58.2	58.2	65.4	65.4	56.4
Grouped	63.6	67.3	70.9	76.4	74.6	69.1

(MFCC, SC, SE) obtained the highest rates when phoneme tokens were used.

4. CONCLUSIONS

Four new subband based features have been proposed for classification of speech under stress. A sequence of simulations has been conducted for different frames sizes, different wavelet packet trees and different speech tokens to find the optimal combination of these free variables to give the highest stress classification scores for both subband based and MFCC based parameters. Simulation results indicate that the new parameters are better suited for stress classification than the MFCC based parameters according to both average scores and the individual best scores. In particular the average score of the subband based cepstrum (SC) parameters achieved +8.6% and +13.6% higher scores than the best of the MFCC based features for grouped and ungrouped stress classification simulations. The best performance of SC on individual simulations achieved +7.3% and +9.1% increases in the classification rates over the best of MFCC based parameters for grouped-stress and ungrouped-stress scenarios, respectively.

ACKNOWLEDGEMENT

The authors wish to thank Dr. John H. L. Hansen for providing the SUSAS database.

References

[1] R. Sarikaya and J. N. Gowdy, "Wavelet Based Analysis of Speech Under Stress," *IEEE Southeascon-97* , pp. 92-96 , April 1997.

[2] R. Sarikaya, "Wavelet Based Classification of Speech Under Stress," M.S Thesis, Clemson University, Clemson, SC, August 1997.

[3] J. H. L. Hansen and B. D. Womack, "Feature Analysis and Neural Network-Based Classification of Speech Under Stress," *IEEE Trans. on Speech and Audio Proc.* , vol. 4, pp. 307-313, 1996.

[4] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with the SUSAS: Speech Under Simulated and Actual Stress Database," *EUROSPEECH-97* , vol. 4, pp. 1743 , 1997.

[5] J. H. L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental robustness in Speech Recognition," *Speech Communication (Special Issue on Speech Under Stress)*, vol. 20, pp. 151-173, 1996.

[6] K. E. Cummings and Mark A. Clements, "Analysis of Glottal Waveforms Across Stress Styles," *ICASSP-90,* pp. 369-372, 1990.

[7] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Communication,* vol. 12, pp. 277-288, 1993.

[8] O. Rioul and M. Vetterli, "Wavelets and Signal Processing," *IEEE Signal Proc. Magazine,* vol. 8(4), pp. 11-38, 1991.

[9] B. J. Stanton, L. H. Jamieson and G. D. Allen, "Acoustic phonetic analysis of loud and lombard speech in simulated cockpit condition," *ICASSP-88,* pp. 331-334, 1988.

[10] J. D. Markel, "Digital Inverse Filtering: A new tool for formant trajectory estimation," *IEEE Transactions on Audio and Electroacoustics,* vol. 20 pp. 129-137, 1972.

[11] E. Erzin, A. E. Cetin and Y. Yardimci, "Subband analysis for speech recognition in the presence of car noise," *ICASSP-95,* vol. 1, pp. 417-420, 1995.

[12] R. J. Schalkoff, "Artificial Neural Networks," *McGraw-Hill,* 1997.